The published version of the paper can be found at:
https://link.springer.com/article/10.1007/s11229-024-04651-7

The Intrapersonal Normative Twin Earth Argument

Abstract:

In this paper I develop an argument against applying a causal theory of mental content to normative concepts. This argument – which I call the Intrapersonal Normative Twin Earth Argument – is inspired by Terry Horgan and Mark Timmons' Moral Twin Earth Argument. The focus of Horgan and Timmons' argument is showing that causal theories of mental content conflict with plausible claims about interpersonal normative disagreement. The Intrapersonal Normative Twin Earth Argument, by contrast, is focused on showing that such theories struggle to vindicate plausible claims concerning whether two of an agent's token normative thoughts have the same or distinct content.

Key Words: Metaethics; Moral Twin Earth Argument; Transparency of Mental Content

1. Introduction

My topic in this paper is the metasemantic question of how normative concepts acquire their semantic content. One answer to this question draws on causal theories of mental content – roughly, views which take as their starting point the idea that the content of a concept is what causes in the right sort of way tokenings of the concept in thought. The classic objection to applying a causal theory of mental content to normative concepts is Terry Horgan and Mark Timmons' Moral Twin Earth Argument (1991; 1992; 1996; 2000; 2008; 2015). In this paper, I develop an argument inspired by the Moral Twin Earth Argument which I call the Intrapersonal Normative Twin Earth Argument. The focus of Horgan and Timmons' argument is showing that causal theories of mental content conflict with plausible claims about interpersonal normative disagreement. The Intrapersonal Normative Twin Earth Argument, by contrast, is focused on showing that such theories struggle to vindicate plausible claims concerning whether two of an agent's token normative thoughts have the same or distinct content.

The value of the Intrapersonal Normative Twin Earth Argument is that it shows that challenges to applying a causal theory of mental content to normative concepts extend beyond capturing disagreement. Also, because of the distinct focus of the argument, it avoids many of the responses that have been developed to the Moral Twin Earth Argument.[1]

The structure of this paper is as follows: in Section (1) I introduce causal theories of mental content and show how such theories might be applied to normative concepts through a discussion of Richard Boyd's causal theory of reference for ethical terms. I also outline the Moral Twin Earth Argument. In Section (2) I develop the Intrapersonal Normative Twin Earth Argument, drawing on work by Paul Boghossian on the transparency of mental content. In Section (3) I defend the Intrapersonal Normative Twin Earth Argument against objections. Section (4) concludes.

2. Causal Theories of Mental Content and the Moral Twin Earth Argument

2.1 Causal Theories of Mental Content

Causal theories of mental content (CTs) hold that the content of at least some concepts is determined by causal relationships between things in the world and mental representations.[2] (Note that proponents of CTs commonly identify concepts with mental representations).[3] To take a prominent example of a CT, Jerry Fodor (1990) claims that a mental representation means, for instance, *cow*, if tokenings of this representation are asymmetrically dependent on cows. Roughly, what this involves is the following: suppose that tokenings of the representation are sometimes caused by things other than cows, say, sheep. This representation nevertheless means *cow* so long as it's the case that if tokenings of the representation weren't caused by cows they wouldn't be caused by sheep and the converse is not the case (1990, pp. 121-122). Putting the theory into possible worlds terms, Fodor tells us that "What's required is just that

[1] Neil Sinhababu (2019) similarly develops objections to applying a causal theory of mental content to normative concepts which are not focused on disagreement. I'm sympathetic to Sinhababu's objections.

[2] For background on CTs see Adams and Aizawa (2021), Rupert (2008), and Neander (2006). These authors observe that causal theories commonly treat the content of mental representations as fundamental and explain the content of items in natural languages in terms of the content of mental representations. Note that everything I say in this paper could be framed at the level of a causal theory of meaning for normative terms. See Devitt and Sterelny (1999, Ch. 4-5) for discussion of a causal theory of meaning at the level of language.

[3] See, for example, Fodor (1998, p. 23) and Rupert (2008, p. 353). In an influential discussion of the ontology of concepts, Eric Margolis and Stephen Laurence similarly hold that "concepts should be identified with mental representations" (2007, p. 588). For worries about this view of concepts see Glock (2009).

there be worlds where cows cause 'cows' and noncows don't; and that they be nearer to our world than any world in which some noncows cause 'cows' and no cows do" (1990, p. 113).

2.2 Boyd on Moral Terms

To see how one might apply a CT to normative concepts, I'll introduce Richard Boyd's attempt to apply a CT to moral terms.[4] Boyd (1988, p. 195) holds that "*Roughly*, and for nondegenerate cases, a term *t* refers to a kind (property, relation, etc.) *k* just in case there exist causal mechanisms whose tendency is to bring it about, over time, that what is predicated of the term *t* will be approximately true of *k.*" Put more simply, a term *t* refers to a property P just in case there is a tendency for our *t* beliefs to get truer of P over time due to our interactions with P. (Notice that Boyd's theory is framed in terms of reference not meaning or content. I assume Boyd is referentialist about content, holding that the content of a term or concept is exhausted by its referential content.)[5]

Applying this theory to the moral domain, Boyd suggests that our use of the term 'good'[6] might be causally regulated in the appropriate way by a 'homeostatic property cluster' consisting of things which satisfy various human needs like love, friendship, health, control over one's own life, physical recreation, intellectual and artistic appreciation, etc. So, the properties that compose goodness are presumably things like being loved, being physically healthy, having friendships, being autonomous, and so on (1988, p. 203).

2.3 The Moral Twin Earth Argument

Turning to the Moral Twin Earth Argument (MTEA), Horgan and Timmons target the MTEA at Boyd's view, which they characterize as committed to the position that "Each moral term t rigidly designates the natural property N that uniquely causally regulates the use of t by humans" (2015, p. 357). Horgan and Timmons invite us to consider a community on a distant planet, Moral Twin Earth, who use a word orthographically and phonetically identical with the English word 'right'. In fact, we can suppose that their entire language is orthographically and

[4] In this paper I'll move freely between claims about language and claims about concepts. For some worries about this (very common) practice see Sawyer (2020).
[5] For defences of referentialism see Fodor (2008, Ch. 3) and Braun (2016). For a discussion of the way in which classic referentialist arguments at the level of language can be applied to the content of concepts see Edwards (2014). For worries about referentialism see Chalmers (2011; 2016) and Segal (2000; 2009).
[6] Note that Boyd is clear that he is specifically discussing moral goodness (1988, p. 203).

phonetically identical to English. Crucially, this community uses their word 'right' in a similar way to the way we use our word 'right'. For example, they apply the term to actions and institutions, they use the term to reason about considerations bearing on well-being, they are normally disposed to act in accordance with their judgements about what is 'right', and they take whether some action falls under the term to be especially important when deciding what to do (Horgan and Timmons, 1991, p. 459; 1992, p. 164). We can add that they commonly use 'right' to commend actions and attitudes and are disposed to feel attitudes of guilt and shame when they perform actions to which they apply their term 'wrong' (Rubin, 2014a, p. 288 & p. 295; 2015a, p. 391; Williams, 2018, p. 42; Horgan and Timmons, 2015, p. 365). Also, they take whether something falls under 'right' (or 'wrong') to supervene on facts they would describe as 'non-moral' (Van Roojen, 2006, p. 172). Now suppose that twin earthlings' use of 'right' is causally regulated in the way spelled out by Boyd's theory by a different property to the property that causally regulates our use of the word 'right'. For instance, perhaps their use is regulated by a property that features in a deontological theory of rightness while our use is regulated by a property which features in a consequentialist theory of rightness (Horgan and Timmons, 2015, p. 358). Horgan and Timmons observe that, if Boyd's theory were true, the truth conditions of my claim that "*x* is right" would be different to the truth conditions of a twin earthling's claim that "*x* is right". The upshot, they suggest, is that in a case in which I say that "*x* is right" and a twin earthling says, "It's not the case *x* is right", we do not disagree with one another; we do not utter sentences with inconsistent contents. However, they suggest that "reflection on this scenario prompts the intuition that the earthling and twin earthling are engaged in a substantive disagreement; they are not simply talking past each other" (2015, p. 359). They summarize the argument as follows: "Boydian moral semantics, if correct as an account of the semantics of terms like 'right', ought to prompt in competent speakers the judgement that the earthling and the twin earthling are not engaged in a genuine moral dispute…rather the dispute is merely verbal. But reflection on the Moral Twin Earth scenario prompts the judgement that the parties are engaged in a genuine moral disagreement" (2015, p. 359).

Horgan and Timmons' argument is often interpreted as supporting the view that the content of moral concepts is (at least in part) determined by their conceptual role, specifically, their 'practical' role in thought – i.e., their connection with action-guiding or motivational

states like intentions or emotions (Rubin, 2014a, pp. 294-296).[7] The argument is also used as evidence for a thesis concerning moral concepts labelled 'referential stability'. According to the referential stability thesis, necessarily, if an agent has a concept M that plays practical role R, then M denotes property P (Williams, 2018, p. 44).[8]

While Horgan and Timmons develop the MTEA with a focus on moral terms or concepts, the argument can straightforwardly be modified to challenge a causal metasemantics for normative concepts (Dunaway and McPherson, 2016, p. 661 & p. 661, footnote 25; Rubin, 2014b, pp. 36-42). While I'll continue to discuss 'The Moral Twin Earth Argument', what I mean by the MTEA from this point onwards (unless explicitly stated otherwise) is the argument modified to focus on normative concepts. By 'normative concepts' I mean the concepts expressed by 'ought' and 'reason' when these terms are used in the sense in which they feature in deliberation, advice, and criticism.[9] I assume that moral concepts are connected to normative concepts. For instance, if something falls under the concept *wrong*, then you have a reason not to do it. (A stronger connection between moral and normative concepts that I won't take a stand on here is that it is a conceptual truth that if something is wrong then you have a reason not to do it (Darwall, 2016).) We can sensibly talk about what an agent morally ought to do using the relevant sense of 'ought'. Such claims can be understood as saying something like the action would be what you ought do if only the distinctively moral reasons were in play (Brown, 2023, p. 10).

3. The Intrapersonal Normative Twin Earth Argument

In this section I'll develop an argument against applying a CT to normative concepts which I call the Intrapersonal Normative Twin Earth Argument (INTEA). Like the MTEA, the INTEA

[7] Also relevant is Eklund (2017, pp. 33-38). For an attempt to develop a cognitivist conceptual role semantics for normative concepts see Wedgwood (2001; 2007, Ch. 4-5). For an argument that a conceptual role semantics for normative concepts is best developed in a non-cognitivist direction see Sinclair (2018).

[8] See also Matti Eklund's discussion of whether there could be 'alternative normative concepts' – concepts that play the same role in thought as our normative concepts, but which differ in reference (2017, Ch. 1-3). Eklund suggests that the best way to reject this possibility is to embrace something like the referential stability thesis. Cf. Boghossian (2021, pp. 379-382).

[9] The word 'ought' is very plausibly a context-sensitive term in the sense that the contribution it makes to the proposition expressed by a sentence in which it features varies depending on the context. However, contextualism about 'ought' is consistent with the claim that we can isolate an 'ought' which is used to express a concept specially connected to deliberation, advice, and criticism. For relevant discussion see Brown (2023), Wedgwood (2018; 2007, Ch. 4-5), Dunaway and McPherson (2016, pp. 642-643), and Hambly (2023). We can similarly isolate a concept expressed by 'reason' that stands in a close connection to the relevant sense of 'ought'; what you ought to do in the relevant sense is determined by what you have reason to do.

draws conclusions about normative metasemantics from reflection on a Twin Earth scenario. However, unlike the MTEA, the INTEA is not concerned with disagreement. Rather, the focus of the argument is on the implications of CTs for intrapersonal sameness and difference of thought content.

3.1 Boghossian on Switching Cases

To develop the INTEA, I'll draw on an argument that Paul Boghossian (1994; 2011; 2015) uses to raise a challenge to externalist views of mental content – i.e., views that hold that the contents of a subject's thoughts are, at least in part, determined by facts which are external to the subject (Farkas, 2008a, p. 326-327; Wikforss, 2008a, p. 161). Accordingly, two subjects who are internal duplicates can have mental states which differ in content.[10] CTs are paradigmatic examples of externalist theories because of the way that they make the content of our thoughts depend on environmental factors.

Boghossian suggests that externalists (or, more precisely, proponents of common forms of externalism inspired by Putnam (1975) and Burge (1979)) are committed to denying a thesis which he calls 'the transparency of mental content': the view that "(a) if two of a thinker's token thoughts possess the same content, then the thinker must be able to know *a priori* that they do; and (b) if two of a thinker's token thoughts possess distinct contents, then the thinker must be able to know *a priori* that they do" (1994, p. 36). By '*a priori*' here Boghossian means independent of 'outer experience' – knowledge based on introspection and memory count as *a priori* in the sense in which he is using the term (1994, p. 33; 2011, p. 475, footnote 1).

One case that Boghossian uses both to illustrate the way that externalist views conflict with the transparency thesis and the cost of denying the thesis is a 'switch case', which involves a subject, call him Switched Peter, who is switched between Earth and Twin Earth. Twin Earth here is Hilary Putnam's (1975, pp. 139-140) Twin Earth: a planet in a distant part of our universe where the stuff that has the same observable features as our water is composed of XYZ rather than H_2O and which, apart from this fact and whatever it entails, is an exact duplicate of Earth. *Twater* is the concept expressed by twin earthlings' uses of 'water', which

[10] Saying what makes two subjects internal duplicates is complicated. For an interesting recent proposal which suggests that two subjects are internally the same just in case they have introspectively indiscriminable mental states see Anil Gomes and Matthew Parrott (2021). To illustrate, consider a Cartesian evil demon scenario. Your mental states in the actual world and in an evil demon scenario are introspectively indiscriminable.

they apply to the stuff composed of XYZ. According to externalists (proponents of CTs among them), the content of this concept is distinct from the content of the earth concept *water*. The classic externalist argument for this conclusion is that the belief that I would express with the sentence "That glass contains water" is true just in case the glass contains H_2O, while the belief my internal duplicate on Twin Earth would express with "That glass contains water" (pointing to the same glass) is true just in case the glass contains XYZ. This difference in truth conditions between the two beliefs is supposed to justify the conclusion that the beliefs (and their constituent concepts) differ in content (Rowlands, Lau, & Deutsch, 2020, §2).[11]

Boghossian (2015, p. 99) contends that, assuming the truth of externalism, for Switched Peter a thesis he calls 'cohabitation' will be true: "earthly and twearthly concepts will commingle in Peter's psychology, without his being aware of that fact. Peter will have both the *water* concept and the *twater* concept, but he will be unaware that he has two 'water' concepts instead of one. Assuming him to be on Twin Earth, and to put it simplistically for present purposes, the *water* concept will get activated when he is recalling 'water' experiences had while on Earth, whereas the *twater* concept will get activated when he is thinking about his current environment."[12] Now suppose that Peter reasons in a way he would express in language as follows:

1) There is water in this lake.

2) There was water in that lake (recalling an experience on Earth).

Therefore,

3) Therefore, there is water in both lakes.

Peter's reasoning commits the fallacy of equivocation; the second premise features the concept *water* while the first features *twater* (2011, p. 459).[13] Boghossian suggests that the conclusion that Peter equivocates is problematic because we have a case in which a rational agent can't,

[11] For internalist responses to this argument see, for instance, Jackson (2003) and Farkas (2008b, Ch. 7). Jackson and Farkas point out that recognizing a difference in truth-conditions between these beliefs is consistent with their content being determined internalistically – so long as it's the case that content determines truth-conditions in combination with facts about the environment occupied by the thinker.

[12] The truth of the cohabitation thesis has been disputed by some philosophers. For discussion see Brown (2004, pp. 170-176). See Boghossian (1994, p. 38) for some points in favour of the thesis.

[13] Some philosophers who accept the cohabitation thesis deny that Switched Peter would equivocate. For instance, Tyler Burge (2013, p. 101) claims that a subject holds "constant, through preservative memory within the argument, the concept used in the first premise in her thinking the second premise." A convincing response to this argument – which points out that it entails that which concept Switched Peter uses will depend on the order in which he thinks of the premises – is developed in Brown (2004, pp. 176-179).

purely introspectively, avoid committing a simple logical fallacy in their reasoning. However, rational agents are able, purely introspectively, to avoid simple fallacious inferences. Boghossian (2015, p. 103) suggests that this claim about rationality is supported by reflection on the fact that irrationality is not a matter of lacking empirical information but rather consists in mishandling the information that one has.

While I think that Boghossian has identified an issue for externalist theories, I'm sympathetic to Åsa Wikforss' (2008b; 2015) suggestion that a more foundational issue for externalism raised by cases like Switched Peter concerns whether content ascriptions explain a subject's cognitive perspective – i.e., explain the way that they reason and act. Externalist views have the implication that an agent might reason and act as if two thoughts feature the same concept when they do not (as the Switched Peter case illustrates) or as if two thoughts feature a different concept when they do not (a possibility illustrated by cases found in Boghossian (1994, p. 37) and Segal (2000, Ch. 3)). Wikforss (2015, pp. 157-162) argues, persuasively to my mind, that the worry that externalist approaches to content fail to capture a subject's cognitive perspective can be developed regardless of whether agents have the sort of meta-beliefs about their own thoughts required by the transparency thesis. Wikforss (2015, p. 162) emphasises that the worry can't be addressed by saying that subjects like Switched Peter don't understand their thoughts and so can't be blamed for reasoning the way they do, observing that "on the externalist view it follows that Peter does not know that he is reasoning invalidly, and this may absolve Peter from his irrationality: he is not to be criticized. However, what the internalist is asking for is not absolution, but an account of content that serves to capture how Peter reasons. And the assertion that he does not know how he reasons does not meet this demand."[14][15]

3.2 Switching Cases and Normative Concepts

My interest in the Switched Peter case is that we can construct a normative analogue of the case to target a CT for normative concepts. Imagine a subject, call him Switched Simon, is

[14] You might think there is an easy solution to the problem Wikforss is raising here: we can simply posit that Switched Peter is presupposing that twater is water. This explains why he reasons as he does. Boghossian (2011, p. 464) points out that this proposal appears to require that Switched Peter could entertain the proposition that <twater is water>. However, he notes that "it's entirely unclear that Peter would understand this proposition in his current state". See also the discussion in Boghossian (2015, pp. 106-107).

[15] To respond to Wikforss' argument, externalists might appeal to various strategies used by proponents of referentialism to reply to cognitive significance-based challenges to their view (raised by Frege's puzzle). For an overview of such strategies and references see Gray (2020, pp. 113-114 & pp. 116-120).

switched between Earth and Normative Twin Earth. Normative Twin Earth is a place where the inhabitants have a concept expressed in their language using 'ought' (call it *tought*) which plays the same role in thought as the concept we express with our word 'ought', but which is causally regulated by a different property to the property that causally regulates the Earth concept. (What is the relevant role in thought? We can, I think, identify this role by examining how the concept features in deliberation, advice, and criticism.)[16] Assuming the truth of a causal theory of mental content for normative concepts, for a subject like Switched Simon, a thesis of cohabitation will be true: earthly and twin earthly concepts will commingle in his psychology. Consequently, Simon will have two 'ought' concepts which differ in content but will be unaware he has two concepts instead of one. Assuming him to be on Normative Twin Earth, the *ought* concept will get activated when he is recalling experiences had while on Earth, whereas the *tought* concept will get activated when he is thinking about his current environment. Suppose that Switched Simon reasons in a way he would express as follows:

1) James did what he ought to do by φ-ing.

2) John did what he ought to by ψ-ing (recalling an experience on Earth).

Therefore,

3) James and John have each done what they ought to do at least once.

Simon's inference commits the fallacy of equivocation; the second premise features the concept *ought* while the first features *tought*. However, I take it to be highly plausible that Switched Simon does not equivocate. His 'ought' term does not change meaning across the premises. But, if Switched Simon doesn't equivocate, then normative concepts don't function as proponents of a CT for normative concepts suggest; *tought* and *ought* do not have distinct contents, even though, by hypothesis, they are causally regulated by different properties. This is the Intrapersonal Normative Twin Earth Argument (INTEA).

3.3 Clarifying the Intrapersonal Normative Twin Earth Argument

In the remainder of Section (2) I'll clarify the INTEA by considering its relationship to Boghossian and Wikforss' objections to externalism which draw on the Switched Peter case and comment on the relationship between the INTEA and the MTEA.

[16] See footnote 9 above for relevant discussion and references.

The INTEA, unlike Boghossian and Wikforss' arguments, appeals directly to the implausibility of the claim that the switched agent equivocates. Why not similarly argue against externalism in the case of the concept *water* by arguing that externalism about *water* implies that (1) Switched Peter equivocates; (2) Switched Peter doesn't equivocate; so, (3) we should reject externalism about *water*? The problem is that it's less clear that Switched Peter does not equivocate than that Switched Simon does not equivocate. I can offer a diagnosis of this difference. Its plausible that if our environment had turned out to be like the Twin Earth environment then the concept we express with 'water' would refer to XYZ not H_2O (Chalmers 2003, pp. 58-59; Jackson, 2003) and *vice versa* for the concept that the twin earthling's express with 'water'. If this is right, Switched Peter's thoughts shift reference across the premises of his argument (on the assumption that one premise features the Earth concept and the other features the Twin Earth concept).[17] On the other hand, it's considerably less plausible that the referent of the concept we express with 'ought' (or the concept expressed by the normative twin earthlings with their use of 'ought') is similarly sensitive to environmental changes. To see this, reflect on the way in which, while knowing whether one is from Earth or Twin Earth appears to be necessary for learning what one's 'water' thoughts refer to, it's not clear that you need to know whether you are from Earth or Normative Twin Earth to determine what (if anything) your 'ought' thoughts refer to (Sinhababu, 2019, pp. 2-5).[18]

It should not be surprising that *water* and *ought* differ with respect to their referential behaviour. What we appear to learn from Putnam's work is that the referent of *water* is determined by the nature of the stuff we were in contact with when the term 'water' was introduced (Putnam, 1975, pp. 141-142; 1990, p. 60; cf. Soames, 2021, p. 89). The idea that normative concepts function in an analogous way is not particularly plausible. When we use 'ought' (and so express the concept *ought*) what we pick out isn't determined by whatever property we were in contact with when the term was introduced. For one thing, this would imply that our ancestors couldn't initially have been wrong about what they ought to do, in much the same way that my parents couldn't initially have been wrong about who is Jesse

[17] Note that I'm not committing myself to the view that a difference in reference entails a difference in content. It's just that what makes the idea that our 'water' concept and the twin earthlings 'water' concept differ in content *prima facie* plausible is that they differ in reference. Cf. footnote 11 above.

[18] Also relevant here are Hattiangadi (2018, p. 605) and Chappell (unpublished manuscript). Hattiangadi and Chappell suggest that normative concepts are plausibly 'super-rigid' in the sense that they have the same referent in all metaphysically possible worlds, and in all epistemically possible scenarios - i.e., centred worlds considered as actual ways the world could be for all we know *a priori*. (For background on the two-dimensional semantic framework Hattiangadi and Chappell are employing see Chalmers (2010).) This feature of normative concepts distinguishes them from a concept like *water* which is such that, while it picks out H_2O in all metaphysically possible worlds, it has a different referent across worlds considered as actual.

Hambly (Pigden, 2012, pp. 104-105; Gampel, 1997, pp. 159-160; cf. Brink, 2001, pp. 163-164).

To be clear about the dialectic, the points I've appealed to in the previous two paragraphs obviously won't be accepted by a proponent of a CT for normative concepts. I discuss these points to explain why, if the independent argument against a CT for normative concepts I've developed succeeds, it doesn't straightforwardly extend to challenge externalism about *water*.

Moving on to the relationship between the INTEA and the MTEA, because of the distinct focus of the INTEA, many challenges to the MTEA do not get any purchase against the INTEA. For instance, attempts to address the MTEA by suggesting that the disagreement between ourselves and the twin earthlings is a matter of metalinguistic disagreement about which concepts to use (Plunket and Sundell, 2013, pp. 19-22) or disagreement in (non-cognitive) attitude (Merli, 2002, pp. 231-239) don't help one address the INTEA. These responses to the MTEA are focused on explaining how there can be disagreement between ourselves and the twin earthlings in a way that doesn't require that we and the twin earthlings are uttering sentences with inconsistent contents.[19] However, even if such non-content-based explanations of normative disagreement succeed, this doesn't bear on the INTEA because the INTEA is not concerned with whether Switched Simon disagrees with anyone (including his former, pre-switch, self). Rather, the INTEA concerns whether Switched Simon's thoughts differ in content across the premises of his argument.

Similarly, consider J.L. Dowell's (2016) influential response to the MTEA. Dowell contends that the probative value of the judgement that we would disagree on moral matters with moral twin earthlings depends on the truth a thesis she calls 'semantic intentionalism'. This is the thesis that "Competence with our moral terms in English requires knowledge of which cross-linguistic similarities in use between our terms and those of any rival, hypothetical language, L', make for sameness of meaning and so the possibility of using our moral terms to express cross-linguistic disagreement with speakers of L'" (2016, p. 11). Dowell argues that semantic intentionalism rests on an implausible view about what is involved in semantic competence, contending that semantic competence with moral (and other) terms requires only an ability to coordinate, communicate, and collect information using such terms (2016, pp. 12-

[19] For discussion of non-content-based approaches to normative disagreement see Finlay (2017).

15).[20] The INTEA does not require that ordinary speakers have the capacity to know what makes for cross-linguistic disagreement with speakers of another language because the INTEA does not make any claims about disagreement (intrapersonal or interpersonal). Rather, the focus of the argument is on intrapersonal sameness of content. Consequently, the argument avoids Dowell's challenge to semantic intentionalism.[21]

4. Defending the Intrapersonal Normative Twin Earth Argument

In this section I'll discuss two objections to the INTEA. The first of these objections is focused on a putative difference between the Switched Peter and Switched Simon cases: In the Switched Peter case, Switched Peter will not be aware that the people around him are using 'water' differently after he is switched. Switched Simon, by contrast, will be aware of a difference with respect to people's use of 'ought'. The second challenge to the INTEA I'll discuss suggests that an influential response to the MTEA which appeals to the phenomenon of reference magnetism can also be used to reply to the INTEA.

4.1 A Disanalogy between Switched Peter and Switched Simon

The first challenge to the INTEA I'll discuss takes as its starting point the claim that Switched Simon would notice that 'ought' is being used differently when transported to Normative Twin Earth. Unlike Switched Peter, who will not be aware of any difference in what falls under the Twin Earth term 'water' and the Earth term 'water' (due to the identical macroscopic appearance of the stuff composed of H_2O and the stuff composed of XYZ), Switched Simon will notice agents applying their 'ought' term differently to the way he applies his 'ought' term. (Notice that this doesn't mean that Simon would realize he had switched planets. He might just

[20] Mark Van Roojen (2018) responds to Dowell by suggesting that the judgement that we disagree with twin earthlings might have probative value even if it isn't supported by our semantic competence with moral terms. Suppose that one is attempting to translate a term 't*' in another language. This doesn't involve drawing on one's semantic competence as Dowell conceives of it (Van Roojen 2018, p. 184). While I think that this is an excellent point, Dowell would probably respond that the people equipped to do translation are field linguists not philosophers (2016, p. 12). I hope that experimental philosophers will take up this issue and poll some field linguists about the Moral Twin Earth Argument or investigate how field linguists go about translating terms as moral terms.

[21] Dowell (2016) also takes aim at a thesis she labels 'intentionalism', very roughly, the view that to be a competent speaker with respect to some term 'a' one must implicitly know what it takes for something to fall into the extension of 'a'. (Dowell notes that Putnam's original Twin Earth Argument arguably presupposes intentionalism.) Dowell's case against intentionalism rests (at least in part) on controversial metasemantic commitments. See footnote 24 below for relevant discussion.

think that, say, people had been subject to widespread brainwashing, or that some other strange event had occurred). For this point to threaten the INTEA, it must be true that (1) being aware of this difference would make it impossible for Simon to acquire the *tought* concept or (2) ensure that Simon would be aware that he had two 'ought' concepts – something which presumably would prevent him from equivocating. In response to (1), surely Simon might join in the normative practices of the normative twin earthlings. After all, it's hardly uncommon for people's normative beliefs on earth to change in line with prevailing community standards. Regarding (2), it's not at all clear that Simon's response to noticing twin earthlings applying 'ought' differently would be to conclude that they are using a different concept. Here on Earth we generally don't think of ourselves as using a different concept to other people when we find that they apply 'ought' differently to us. In fact, one of the main attractions of a causal metasemantics for normative concepts is that it promises a way of making sense of how we here on Earth all could be thinking and talking about the same thing despite differences in what we take to fall into the extension of 'ought' (Boyd, 1988, p. 199; Sayre-McCord, 1997, p. 281). Perhaps if the differences in how twin earthlings applied their 'ought' term were massive then Simon might conclude that they were using a different concept.[22] However, I don't see any reason for thinking that there must be such a large difference between earthlings and twin earthlings use of 'ought' in order for it to be true that the concepts they are expressing are causally regulated by different properties. Perhaps, for example, twin earthlings apply their 'ought' term differently to how earthlings apply their term because they are less concerned with prudential considerations than earthlings (Rubin, 2014b, p. 37-38).

4.2 Reference Magnetism

The next objection to the INTEA I'll consider emerges from an objection to the MTEA. This objection to the MTEA, unlike those I discussed in Section (3.3), does constitute a challenge to the INTEA. Several theorists (Van Roojen, 2006; Edwards, 2013; Dunaway and McPherson, 2016) have appealed to reference magnetism to respond to the MTEA. Reference magnetism is the idea that certain properties – the 'natural' or elite' properties which 'carve reality at its

[22] Of relevance here are discussions of the Moral Twin Earth case which suggest that there are 'substantive constraints' on what can fall under moral concepts. For instance, perhaps we wouldn't recognize a concept as a moral concept unless it tracks considerations of harms and benefits (although cf. Rubin, 2008, pp. 317-318). Rubin (2014a, pp. 296-302) convincingly argues that the MTEA is compatible with the existence of substantive constraints on moral concepts.

joints' – are more eligible candidates for reference than other properties; these properties 'attract' reference (Lewis, 1983, pp. 370-377; 1984, pp. 226-227).

Dunaway and McPherson (2016) illustrate how reference magnetism might be used to respond to the MTEA by imagining a metasemantic theory (call it the 'Toy Theory') according to which the referent of 'ought' is determined by the reference assignment that maximizes the sum of (1) fit with use, understood in terms of the extent to which it makes sentences featuring 'ought' which are accepted by the community of speakers come out as true, and (2) the eliteness of a candidate referent. Now suppose that there is a single highly elite property in the vicinity of our use of 'ought' and the twin earthlings use of 'ought' (which Dunaway and McPherson label 'ought*'). Dunaway and McPherson suggest that, "Here, thanks to reference magnetism, the Toy Theory of reference suggests that we and our twins refer to the same property with our use of the words 'ought' and 'ought*' respectively. In other words, reference magnetism can vindicate the core semantic judgment that is the heart of the Normative Twin Earth challenge" (2016, p. 666).

Suppose we build reference magnetism into a CT – for concreteness, I'll focus on Boyd's view.[23] Assume that properties can do better or worse at meeting the causal-epistemic constraint on reference determination central to Boyd's theory according to which term *t* refers to property P just in case there is a tendency for our *t* beliefs to get truer of P over time due to our interactions with P. (This is something which has been doubted (Edwards, 2013, p. 12), but it strikes me as plausible.) Now suppose that there is some property on Earth and Twin Earth which both does well meeting Boyd's causal-epistemic constraint and is highly elite. Perhaps we and the twin earthlings both refer to this elite property with our respective uses of 'ought'.

The relevance of this discussion for the INTEA should be clear: one might argue that, with reference magnetism added to the theory, Boyd's view (or some other CT) won't have the consequence that Switched Simon acquires a concept with distinct content when transported to Twin Earth because *ought* and *tought* are in fact co-referential.

[23] Building reference magnetism into Boyd's theory is something that theorists have suggested needs to be done for reasons unrelated to addressing the MTEA. Suikkanen (2017, pp. 10-13) contends that adding reference magnetism to the theory is needed to deal with the 'qua-problem' for causal theories of reference. Roughly, this is the problem that whenever we are in causal contact with some object (property or entity) there will be a multitude of other objects which we are also in causal contact with (Devitt and Sterelny 1999, pp. 79-82 & pp. 90-93). For instance, suppose you are in perceptual contact with the property of being green. You are also in perceptual contact with the property of being grue. Zhao (2021, pp. 11165-11166) also stresses the importance of reference magnetism for helping to tackle the worry that Boyd's theory fails to fix determinate reference.

My primary response to the challenge to the INTEA (and MTEA) from the appeal to reference magnetism will unfold through an examination of the notion of eliteness. I'll argue that it is unclear that there is an interpretation of eliteness that is both plausible when applied to normative properties and helpful to the proponent of a CT for normative concepts without requiring serious revisions to the view. However, before I develop this response, I want to register some general reservations about reference magnetism. What I take to be a particularly significant worry comes from reflecting on cases like the following. Suppose that speakers apply a term 't' to all and only those things which they take to be F and that hearers take utterances of 't' as evidence that something is F. However, due to reference magnetism, 't' refers to some other more natural property G. The problem here is the following: we have a situation where what we communicate with our use of 't' comes apart from the referent of 't' (Schwarz, 2016, pp. 14-18; Cohnitz and Haukioja, 2020, pp. 128-135).[24]

Putting aside these reservations and turning to the question of how to understand eliteness, the first way of cashing out the notion of eliteness I'll consider is drawn from David Lewis's (1983; 1984; 1986) seminal discussion of eliteness.[25] Lewis suggests that there are perfectly elite properties – properties which make for objective similarity so that things that share them are qualitative duplicates and which form the supervenience base for all other properties in our world (1983, pp. 355-364). Other properties are more or less elite depending on the length of their definition in terms of the perfectly elite properties (1984, p. 228; 1986, p. 61). Lewis combines these claims about the structural features of eliteness with the suggestion that the perfectly elite properties are the fundamental physical properties (mass, charge, spin, etc.)

[24] Cohnitz and Haukioja (2020; 2016) argue that the sort of worry about reference magnetism I've discussed here can be raised for any view committed to the position that which theory of reference is true for an expression is not determined by the psychological states of speakers. Views of this kind – which they label 'meta-externalist' – are consistent with scenarios where we all use and interpret expression 'e' as if it refers to P but in fact the expression refers to Q. For instance, imagine a world where everyone agreed that 'Gödel' referred to the unique individual, if any, who proved the incompleteness theorem and used 'Gödel' in a way consistent with this conclusion but the expression in fact had a Kripkean (1980, lecture II) semantics. Cf. Jackson (2009, p. 410). Cohnitz and Haukioja point out that Kripke and Putnam are not meta-externalists. Kripke and Putnam argue that external facts are relevant to determining reference facts for certain terms *because* of our intentions with respect to these terms (Putnam, 1975, pp. 141-142; Kripke, 1980, pp. 134-135). J.L. Dowell, in their highly influential discussion of MTEA, seems to adopt a meta-externalist view (2011, p. 18-25). I think that this is in tension with their emphasis on the idea that a semantic theory is to be evaluated in terms of how well it explains (among other things) our capacity to communicate using language.

[25] Lewis's discussion (along with much of the subsequent literature) is framed in terms of 'naturalness' not 'eliteness'. However, I'll follow Dunaway and McPherson (2016, p. 646) in using 'eliteness' given that – as they point out – this terminology is less likely to cause confusion in the metaethical context.

As several theorists have argued (Dunaway and McPherson, 2016, p. 652; Schroeter and Schroeter, 2013, p. 16-19; Rubin, unpublished manuscript)[26] it's not clear that Lewis' approach to eliteness is helpful for tackling the MTEA (or the INTEA). Consider a definition of a property which is a candidate for the property of being what one ought to do, e.g., the property of maximizing net happiness, using only predicates that refer to the properties of fundamental physics. This definition may be infinitely long given that this property may be infinitely physically realizable.[27] But if definitions of candidate properties are infinitely long then such properties are presumably equally (un)natural. (I'm going to ignore a further complication raised by the fact that there are presumably possible worlds where candidate properties are not realized by the fundamental physical properties of our world.) Also, even if definitions of candidate properties are finite, it's not obvious that we are in any position to learn about the relative length of such definitions – at least for some candidate definitions (although cf. Mokriski, 2020, pp. 718-723). However, to my mind, the most significant problem with employing this account of eliteness is more straightforward. As Schroeter and Schroeter (2013, p. 17) put the point, "the idea that the reference of ['what one ought to do'] could depend on how many logical connectives it takes to define a referential candidate in the language of microphysics seems incredible: intuitively, [what one ought to do] is not beholden to microphysics in this way."

These sorts of issues have led philosophers to develop alternative accounts of eliteness. For example, Dunaway and McPherson (2016, pp. 652-653) reject Lewis's view that relative eliteness is a matter of definability in terms of the perfectly elite properties. Instead, they develop a view according to which relative eliteness rather than perfect eliteness is treated as primitive. They also offer an epistemology of relative eliteness, claiming that we come to "know which properties are highly elite by knowing which properties are countenanced by naturalistically credible theoretical disciplines including (but not limited to) physics" (2016, p. 654; cf. Van Roojen, 2006, p. 81). Moreover, they suggest that normative theorizing may count as a naturalistically credible discipline.[28]

---

[26] I'd like to thank Michael Rubin for giving me permission to cite this paper and for his help with the ideas in this sub-section.

[27] While I'll follow the practice in the literature, I want to register that I find the language of 'definition' in this context problematic. As Gideon Rosen points out (2015, pp. 192-193), the idea that a massive disjunction which lists the conditions which necessitate the instantiation of a property is a definition of this property seems wrong. To illustrate, consider 'defining' the property of being a prime number in terms of being a 2 or a 3 or a 5… . .

[28] Of relevance here is Dunaway and McPherson's (2016, pp. 655-656) claim that one can't use their account of eliteness to respond to the MTEA by suggesting that (1) normative properties are reducible to (say) psychological properties and (2) that psychology counts as a naturalistically credible discipline. The problem that they identify

What would it take to establish normative theorizing is a naturalistically credible discipline? One answer suggested by certain elements of Dunaway and McPherson's discussion is that this depends on the explanatory credentials of normative properties.[29] Dunaway and McPherson tell us that normative properties "explain facts about normativity, action-guidingness, and the like" (2016, p. 655). As it stands, this claim is not particularly clear. Michael Rubin (unpublished manuscript) helpfully suggests two disambiguations. According to the first disambiguation, normative properties explain non-normative facts such as facts about agents' beliefs, motivational states, or actions. On the second disambiguation, the natural properties that normative properties reduce to (according to the sort of naturalist normative realism embraced by Dunaway and McPherson)[30] explain normative facts such as the fact that I ought to go to the party tonight.

Consider first the disambiguation according to which the explananda are non-normative facts. Rubin points out that it's hard to see why distinct properties on Normative Twin Earth couldn't play the same explanatory role there. And, if this is true, there won't be a single highly elite property in the vicinity of both our and the twin earthling's use of 'ought', as Dunaway and McPherson's response to the MTEA requires. To illustrate the claim that different properties could play the same explanatory role across the planets, Rubin considers the explanations of non-moral facts by moral properties offered by some proponents of naturalistic moral realism – e.g., that injustice explains social instability (Brink, 1989, pp. 187-197). This explanatory claim looks as though it could be true on Normative Twin Earth even if injustice on Normative Twin Earth is a distinct property to injustice on Earth. While there is some plausibility to the idea that grossly inegalitarian societies are likely to be unstable, societies which are just by, say, Kantian standards or by rule-utilitarian standards are unlikely to display such a grossly inegalitarian character. Justice could be a Kantian property on Earth and a rule-utilitarian property on Normative Twin Earth and the explanation of social instability will work just as well on both planets. The same sort of point Rubin develops can be made with respect to explanations of agents' moral beliefs which cite moral properties – e.g., that Andrew's belief that the boys' action of lighting the cat on fire was wrong is explained by the fact that the boys'

with this strategy is that one could defend the MTEA by finding two communities whose respective uses of 'ought' are governed by distinct but similarly elite psychological properties. Williams (2018, pp. 57-58) suggests that Dunaway and McPherson's own account runs into a somewhat related problem.

[29] I'm not entirely sure that this is the correct interpretation of their view. Some of their comments suggest that they think that whether normative theorizing is a naturalistically credible discipline just depends on whether there is "a naturalistically acceptable epistemology for the normative" (2016, p. 656).

[30] See Dunaway and McPherson (2016, p. 644).

action was wrong. Sturgeon (1986, pp. 246-247) suggests that if the boys' action didn't have certain natural properties, assumed for the sake of argument to be the supervenience base for the wrongness of the action, then Andrew wouldn't have formed the belief it was wrong. But, again, this explanation looks like it will work for Twin Andrew's belief on Normative Twin Earth that the boys' action was wrong – even if wrongness is a distinct property on Normative Twin Earth.

Turning to the second disambiguation, suppose that what is to be explained are normative facts such as the fact that I ought to go to the party. For my purposes in this paper, I'm happy just to note that if appealing to reference magnetism to resist the MTEA and INTEA involves taking this path, we've departed a very long way from a classic causal theory of content. We now have a picture according to which what *ought* refers to is, to a significant extent, determined by which property features in the best normative theory – i.e., the sort of theory that normative ethicists are in the business of developing (cf. Sayre-McCord, 1997).[31] Importantly, the goodness of a normative theory in the relevant sense isn't a matter of the causal explanations it offers. Explanations of normative facts are *not* causal explanations. According to several popular metaethical views such explanations are instead metaphysical explanations; the explanation for why it's true that I ought to go to the party will cite facts which metaphysically explain why I ought to go to the party, e.g., that it will bring me pleasure.[32]

5. Conclusion

In this paper I've argued that, when applied to normative concepts, causal theories of mental content conflict with plausible claims about intrapersonal sameness of normative content. This challenge is distinct from the Moral Twin Earth Argument which is focused on whether such theories have problematic implications concerning normative disagreement. That said, I see the two arguments as complementing each other, helping to build a cumulative case against a causal metasemantics for normative concepts. While I've focused on causal theories of mental

[31] Sayre McCord (1997, p. 291) claims that "what a moral term refers to, if anything, is determined by whether, in light of the best moral theory, the use of that term can be seen as appropriately regulated by instances of a normatively significant kind. Our sincere deployment of [moral] terms reflects…our conviction that we are using them to refer to what the best theory would reveal to be normatively significant kinds."

[32] For discussion of normative explanation see Fogal and Risberg (2020). Notice that I've focused on the explanation of *particular* normative facts – i.e., normative facts concerning dated, non-repeatable things such as an action, person, or state of affairs (Fogal and Risberg, 2020, p. 172) rather than the explanation of general normative facts (normative principles).

content in this paper, the challenge I've developed for such theories will extend to any other metasemantic theory similarly committed to holding that Switched Simon equivocates.

Acknowledgments

I would like to thank Nic Southwood, Shang Long Yeo, Hezki Symonds, Josef Holden, Philip Pettit, Frank Jackson, the members of the Australian National University Philosophy of Mind Work-in-Progress Group, and several anonymous referees for their helpful comments. My greatest thanks go to Michael Rubin for his generous feedback on the ideas in this paper.

Bibliography

Adams, F., & Aizawa, K. (2021). Causal Theories of Mental Content. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, Fall 2021 edition. https://plato.stanford.edu/archives/fall2021/entries/content-causal/.

Boghossian, P. (1994). The Transparency of Mental Content. *Philosophical Perspectives*, 8, 33–50.

Boghossian, P. (2011). The Transparency of Mental Content Revisited. *Philosophical Studies*, 155(3), 457–65.

Boghossian, P. (2015). Further Thoughts on the Transparency of Mental Content. In S. C. Goldberg (Ed.), *Externalism, Self-Knowledge, and Skepticism: New Essays*, 97–112. Cambridge University Press.

Boghossian, P. (2021). Normative Principles Are Synthetic A Priori. *Episteme*, 18(3), 367–83.

Boyd, R.N. (1988). How to be a Moral Realist. In G. Sayre-McCord (Ed.), *Essays on Moral Realism*, 181-229. Cornell University Press.

Braun, D. (2016). The Objects of Belief and Credence. *Mind*, 125(498), 469–97.

Brink, D.O. (2001.) Realism, Naturalism, and Moral Semantics. *Social Philosophy and Policy*, 18(2), 154–76.

Brink, D.O. (1989.) *Moral Realism and the Foundations of Ethics*. Cambridge University Press.

Brown, J. (2004). *Anti-Individualism and Knowledge*. MIT Press.

Brown, J.L.D. (2023). On Scepticism about Ought Simpliciter. *Australasian Journal of Philosophy*. https://doi.org/10.1080/00048402.2023.2225527.

Burge, T. (1979). Individualism and the Mental. *Midwest Studies in Philosophy*, 4(2), 73-121.

Burge, T. (2013). Memory and Self Knowledge. In *Cognition Through Understanding: Self-Knowledge, Interlocution, Reasoning, Reflection: Philosophical Essays*, Volume 3, 88-103. Oxford University Press.

Chalmers, D. (2016). Referentialism and the Objects of Credence: A Reply to Braun. *Mind*, 125(498), 499-510.

Chalmers, D. (2011). Frege's Puzzle and the Objects of Credence. *Mind*, 120(479), 587–635.

Chalmers, D. (2010). "Appendix: Two-Dimensional Semantics." In *The Character of Consciousness*. Oxford University Press.

Chalmers, D. (2003). "The Nature of Narrow Content." *Philosophical Issues*, 13, 46-66.

Cohnitz, D., & Haukioja, J. (2020). Variation in Natural Kind Concepts. In T. Marques & A. Wikforss (Eds.), *Shifting Concepts: The Philosophy and Psychology of Conceptual Variability*, 128-146. Oxford University Press.

Cohnitz, D., & Haukioja, J. (2013). Meta-Externalism vs Meta-Internalism in The Study of Reference. *Australasian Journal of Philosophy*, 91(3), 475–500.

Darwall, S. (2016). Making the 'Hard' Problem of Moral Normativity Easier. In E. Lord & B Maguire (Eds.), *Weighing Reasons*, 257-278. Oxford University Press.

Devitt, M., & Sterelny, K. (1999). *Language and Reality: An Introduction to the Philosophy of Language*, Second Edition. Blackwell.

Dowell, J.L. (2016). The Metaethical Insignificance of Moral Twin Earth. In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics*, Vol. 11, 1-27. Oxford University Press.

Dunaway, B., & McPherson T. (2016). Reference Magnetism as a Solution to the Moral Twin Earth Problem. *Ergo*, 3(25), 639–679.

Edwards, K. (2014). Keeping (Direct) Reference in Mind. *Noûs*, 48(2), 342–67.

Eklund, M. (2017). *Choosing Normative Concepts*. Oxford University Press.

Farkas, K. (2008a). Semantic Internalism and Externalism. In E. Lepore & B. C. Smith (Eds.), *The Oxford Handbook of the Philosophy of Language*, 323-340. Oxford University Press.

Farkas, K. (2008b). *The Subject's Point of View*. Oxford University Press.

Finlay, S. (2017). Disagreement Lost and Found. In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics*, Vol. 12, 187-205. Oxford University Press.

Fodor, J. (2008). *LOT2: The Language of Thought Revisited*. Oxford University Press.

Fodor, J. (1998). *Concepts: Where Cognitive Science Went Wrong*. Oxford University Press.

Fodor, J. (1990). A Theory of Content, II: The Theory. In *A Theory of Content and Other Essays,* 89-136. MIT Press.

Fogal, D., & Risberg, O. (2020). The Metaphysics of Moral Explanations. In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics*, Vol. 15, 170-194. Oxford University Press.

Gampel, E. H. (1997). Ethics, Reference, and Natural Kinds. *Philosophical Papers*, 26(2), 147–163.

Glock, H. (2009). Concepts: Where Subjectivism Goes Wrong. *Philosophy*, 84(1), 5–29.

Gomes, A., & Parrott M. (2021). On Being Internally the Same. In U. Kriegel, *Oxford Studies in Philosophy of Mind*, Vol. 1, 315-340. Oxford University Press.

Gray, A. (2021). Cognitive Significance. In S. Biggs & H. Geirsson*, The Routledge Handbook of Linguistic Reference*, 107-122. Routledge.

Hambly, J. (2023). Advice for Analytic Naturalists. *Ergo*, 9(59), 1604-1626.

Hattiangadi, A. (2018). Moral Supervenience. *Canadian Journal of Philosophy*, 48(3–4), 592–615.

Horgan, T., & Timmons, M. (2015). Exploring Intuitions on Moral Twin Earth: A Reply to Sonderholm. *Theoria*, 81(4), 355–75.

Horgan, T., & Timmons, M. (2008). Analytical Moral Functionalism Meets Moral Twin Earth. In Ian Ravenscroft (Ed.), *Mind, Ethics and Conditionals: Themes from the Philosophy of Frank Jackson*, 221-235. Oxford University Press.

Horgan, T., & Timmons, M. (2000). Copping out on Moral Twin Earth. *Synthese*, 124(1), 139–52.

Horgan, T., & Timmons, M. (1996). From Moral Realism to Moral Relativism in One Easy Step. *Crítica*, 28(83), 3-39.

Horgan, T., & Timmons, M. (1992). Troubles for New Wave Moral Semantics: The 'Open Question Argument' Revived. *Philosophical Papers*, 21(3), 153-175.

Horgan, T., & Timmons, M. (1991). New Wave Moral Realism Meets Moral Twin Earth. *Journal of Philosophical Research*, 16, 447–465.

Jackson, F. (2009). Replies to Critics. In Ian Ravenscroft (Ed.), *Mind, Ethics and Conditionals: Themes from the Philosophy of Frank Jackson*, 387-486. Oxford University Press.

Jackson, F. (2003). Narrow Content and Representation, or Twin Earth Revisited. *Proceedings and Addresses of the American Philosophical Association*, 77(2), 55–70.

Kripke, S. (1980). *Naming and Necessity*. Harvard University Press.

Lewis, D. (1983). New Work for a Theory of Universals. *Australasian Journal of Philosophy*, 61(4), 343–77.

Lewis, D. (1986). *On the Plurality of Worlds*. Blackwell.

Lewis, D (1984). Putnam's Paradox. *Australasian Journal of Philosophy*, 62(3), 221–36.

Margolis, E., & Laurence, S. (2007). The Ontology of Concepts-Abstract Objects or Mental Representations? *Noûs*, 41(4), 561–93.

Merli, D. (2002). Return to Moral Twin Earth. *Canadian Journal of Philosophy*, 32(2), 207–240.

Mokriski, D. (2020). The Methodological Implications of Reference Magnetism on Moral Twin Earth. *Metaphilosophy*, 51(5), 702–26.

Neander, K. (2006). Naturalistic Theories of Reference. In M. Devitt & R. Hanley (Eds.), *The Blackwell Guide to the Philosophy of Language*, 374-391. Blackwell.

Pigden, C. (2012). Identifying Goodness. *Australasian Journal of Philosophy*, 90(1), 93–109.

Plunkett, D. & Sundell, T. (2013). Disagreement and the Semantics of Normative and Evaluative Terms. *Philosopher's Imprint*, 13(23), 1-37.

Putnam, H. (1975). The Meaning of Meaning. *Minnesota Studies in the Philosophy of Science*, 7, 131-193.

Putnam, H. (1990). *Realism with a Human Face*. Harvard University Press.

Rosen, G. (2015). Real Definition. *Analytic Philosophy*, 56(3), 189–209.

Rowlands, M., Lau, J & Deutsch M. (2020). Externalism About the Mind. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*, Winter 2020 Edition. https://plato.stanford.edu/archives/win2020/entries/content-externalism.

Rubin, M. (2015). Normatively Enriched Moral Meta-Semantics. *Philosophy and Phenomenological Research*, 91(2), 386–410.

Rubin, M. (2014a). Biting the Bullet on Moral Twin Earth. *Philosophical Papers*, 43(2), 285–309.

Rubin, M. (2014b). On Two Responses to Moral Twin Earth. *Theoria*, 80(1), 26–43.

Rubin, M. (2008). Sound Intuitions on Moral Twin Earth, *Philosophical Studies*, 139(3), 307–327.

Rubin, M. Unpublished manuscript. Reference Magnetism Offers No Solution to Normative Twin Earth. Available on request from michael.rubin@uwa.edu.au.

Rupert, R. (2008). Causal Theories of Mental Content. *Philosophy Compass*, 3(2), 353–380.

Sawyer, S. Talk and Thought. In A. Burgess, H. Cappelen, & D. Plunkett (Eds.) *Conceptual Ethics and Conceptual Engineering*, 379-395. Oxford University Press.

Sayre-McCord, G. (1997). 'Good' On Twin Earth. *Philosophical Issues*, 8, 267-292.

Schroeter, L., & Schroeter, F. (2013). Normative Realism: Co-Reference Without Convergence? *Philosophers Imprint*, 13(13), 1-24.

Schwarz, W. (2014). Against Magnetism. *Australasian Journal of Philosophy*, 92(1), 17–36.

Segal, G. (2000). *A Slim Book About Narrow Content*. MIT Press.

Segal, G. Keep Making Sense. *Synthese*, 170(2), 275-287.

Sinclair, N. (2018). Conceptual Role Semantics and the Reference of Moral Concepts. *European Journal of Philosophy*, 26(1), 95–121.

Sinhababu, N. (2019). One-Person Moral Twin Earth Cases. *Thought*, 8(1), 16–22.

Soames, S. (2021). Fruits of the Causal Theory of Reference. In S. Biggs & H. Geirsson (Eds.), *The Routledge Handbook of Linguistic Reference*, 82-93. Routledge.

Sturgeon, N. (1988). Moral Explanations. In G. Sayre-McCord (Ed.), *Essays on Moral Realism*, 229-255. Cornell University Press.

Suikkanen, J. (2017). Non-Naturalism and Reference. *Journal of Ethics and Social Philosophy*, 11(2), 1-24.

van Roojen, M. (2018). Rationalist Metaphysics, Semantics and Metasemantics. In K. Jones & F. Schroeter (Eds.), *The Many Moral Rationalisms*, 167-186. Oxford University Press.

van Roojen, M. (2006). Knowing Enough to Disagree: A New Response to the Moral Twin Earth Argument. In R. Shafer-Landua (Ed.), *Oxford Studies in Metaethics*, Vol. 1, 161-194. Oxford University Press.

Wedgwood, R. (2018). The Unity of Normativity. In D. Star (Ed.), *The Oxford Handbook of Reasons and Normativity,* 23-45. Oxford University Press.

Wedgwood, R. (2007). *The Nature of Normativity*. Oxford University Press.

Wedgwood, R. (2001). Conceptual Role Semantics for Moral Terms. *The Philosophical Review*, 110(1), 1–30.

Wikforss, A. M. (2015). The Insignificance of Transparency. In S. C. Goldberg (Ed.), *Externalism, Self-Knowledge, and Skepticism: New Essays*, 142–164. Cambridge University Press.

Wikforss, A. M. (2008a). Semantic Externalism and Psychological Externalism. *Philosophy Compass*, 3(1), 158-181.

Wikforss, A. M. (2008b). Self-Knowledge and Knowledge of Content." *Canadian Journal of Philosophy*, 38(3): 399–424.

Williams, J. R. G. 2018. "Normative Reference Magnets." *The Philosophical Review*, 127(1): 41–71.

Chappell, R.Y. Unpublished manuscript. "The 2-D Argument Against Metaethical Naturalism." Available from https://philpapers.org/rec/CHAMSA-2. Accessed August 19, 2023.

Zhao, Xinkan. 2021. “Metasemantics and Boydian Synthetic Moral Naturalism.” *Synthese*, 199 (3–4): 11161–11178.